

计算机

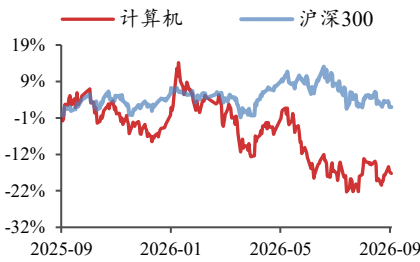
2026 年 09 月 06 日

高密度算力驱动液冷加速渗透，AI 产业持续迭代

——行业周报

投资评级：看好（维持）

行业走势图



数据来源：聚源

相关研究报告

《算力供需缺口持续扩大，算力租赁行业景气度上行——行业深度报告》
-2026.8.25

《DeepSeek API 拟大幅提价，算力景气度高企——行业点评报告》-2026.8.9

《DeepSeek-V4-Flash 超高性价比重塑 Agent 竞争格局——行业点评报告》
-2026.8.3

张越（分析师）

zhangyue1@kysec.cn

证书编号：S0790524090003

叶进新（联系人）

yejinxin@kysec.cn

证书编号：S0790126060053

● 液冷由可选走向标配，高密算力持续打开产业成长空间

AI 服务器功耗与单机柜功率密度持续提升，液冷已从传统辅助散热升级为高密算力基础设施的确定性刚需。当前冷板式液冷凭借技术成熟、供应链完善及对现有服务器架构的兼容性占据主流；下一代 AI 芯片功耗进一步提升，两相冷板式将成为高功率密度场景的重要技术储备，浸没式则持续探索超高密度场景。液冷价值集中于二次侧，CDU、冷板等核心部件壁垒与价值占比较高，产业价值正向核心零部件集中。同时液冷已从服务器配套升级为影响算力释放、能耗与交付效率的核心基础设施。国内液冷产业体系加快完善，多类厂商持续加码，竞争正从单一部件供应向系统集成和全栈散热能力升级。

● 大模型、Agent 与国产算力多线共振，AI 产业生态持续完善

本周 AI 产业延续高频迭代，模型、Agent、国产及海外算力均有新进展。模型端，DeepSeek-V4 实验性多模态开源，腾讯 WorkBuddy 开放平台上线并开放 Agent 底层能力，厂商持续优化模型性能、效率与成本，AI 能力向 Agent 平台和产业生态延伸。国产算力侧，长鑫 LPDDR6 量产，华虹宏力获增资，范式智能拟采购华为昇腾芯片，正从技术突破向规模化应用推进。海外厂商持续加码算力基建。当前产业仍处模型、应用、算力互促阶段，国产适配深化，有望推动国内 AI 产业链自主化与商业化。

● 计算机板块跑赢主要指数，AI 应用表现相对占优

8 月 31 日至 9 月 4 日，计算机指数上涨 0.49%，同期沪深 300、上证指数、创业板指和科创 50 分别下跌 1.33%、0.56%、4.03%和 5.10%，计算机板块分别跑赢 1.82、1.05、4.52 和 5.59 个百分点，在 31 个一级行业中排名第 12 位，整体展现出较强的相对韧性。AI 细分方向表现分化，AI 应用指数全周累计上涨约 1.81%，明显优于 AI 算力指数；个股层面，AI 应用板块涨幅前十个股周涨幅均超过 11%，赚钱效应相对突出，AI 算力板块虽指数小幅承压，但内部仍存在明显的结构性机会。

● 受益标的

国产算力：寒武纪、海光信息、禾盛新材、沐曦股份、浪潮信息、工业富联等；
海外算力：东山精密、胜宏科技、联特科技等；
算力租赁：宏景科技、利通电子、协创数据等；
大模型：智谱、MiniMax、阿里巴巴、腾讯控股；
液冷：同飞股份、恒铭达等。

● **风险提示：**大模型技术迭代不及预期风险、AI 技术应用不及预期风险、相关政策不及预期风险、AI Infra 建设不及预期风险等。

目 录

1、液冷：AI 时代下高密算力的确定性刚需	3
1.1、AI 服务器算力密度跃升，液冷迈入标配时代	3
1.2、二次侧价值量占比较高，核心部件贡献主要份额	8
1.3、智算液冷市场持续扩张，国内产业体系加速成型	13
2、计算机产业动态更新	14
2.1、CSP 厂商与大模型厂商最新动向	14
2.2、国产算力产业最新进展	16
2.3、海外算力产业最新进展	18
3、上周市场表现回顾（2026/8/31-2026/9/4）	19
4、风险提示	21

图表目录

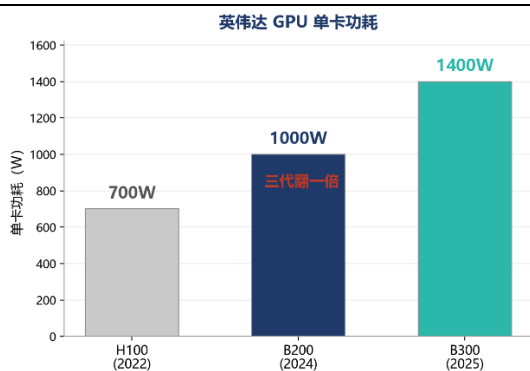
图 1：从英伟达 H100 到 B300，单卡功耗翻倍增长	3
图 2：传统风冷散热极限约在 30kW/柜，已触及物理极限	3
图 3：风冷数据中心已无法满足最新 PUE 要求，液冷成为唯一可行的技术路径	3
图 4：冷板式液冷采用微通道强化换热技术具有极高的散热性能	4
图 5：Vera Rubin NVL72 采用 45°C 单相温水直冷，降低冷水机组依赖与制冷能耗	4
图 6：两相冷板式冷却效率更好，但其对系统元件介质兼容性、管道承压等要求更高	5
图 7：液态工质吸热蒸发为气液两相，靠相变潜热带走芯片热量	6
图 8：两相冷板式液冷系统采取双回路设计，覆盖 ICT 冷板、CDU 单元与室外冷却塔	6
图 9：传热介质在热量传递过程中仅发生温度变化，而不存在相态转变	7
图 10：传热介质在热量传递过程中会发生相态转变，依靠物质的潜热变化传递热量	7
图 11：喷淋式液冷实现了 100% 液冷，其结构颠覆性优于浸没式液冷	8
图 12：液冷系统价值量分布主要集中于二次侧	8
图 13：核心部件贡献液冷系统主要价值量份额	8
图 14：CDU 泵送冷却液冷却服务器，回流至外部冷源散热，形成闭环冷却循环	9
图 15：冷却液经 CDU 送入芯片冷板吸热，回流至散热端冷却循环	10
图 16：QDC 是冷板式液冷和浸没式液冷循环中唯一的动态连接界面	10
图 17：UQDC 为盲插型 QDC，具备高容差补偿，支持盲插，适配高密机架运维	11
图 18：Manifold 是连接液冷机柜 CDU 与冷板之间主管路的关键部件	12
图 19：英伟达 Rubin 高密机架定制升级液冷歧管	12
图 20：英伟达 MGX 44RU 机架通用标准液冷歧管	12
图 21：计算机行业板块相对表现	20
图 22：AI 细分指数表现情况	20
图 23：AI 算力板块个股本周涨幅前十	21
图 24：AI 应用板块个股本周涨幅前十	21
表 1：两相冷板式液冷有望成为下一代高功率机柜储备方案	5
表 2：QDC 设计涉及流体力学、材料科学、精密加工及可靠性工程等多个交叉学科	11
表 3：冷却液是液冷系统完成热量传递的核心工质，其选型须依据具体应用场景，并满足多维度性能要求	13
表 4：AI 服务器液冷产业链呈上、中、下游清晰分层，产业竞争正由单一部件角逐向全栈散热能力升级	13

1、液冷：AI 时代下高密算力的确定性刚需

1.1、AI 服务器算力密度跃升，液冷迈入标配时代

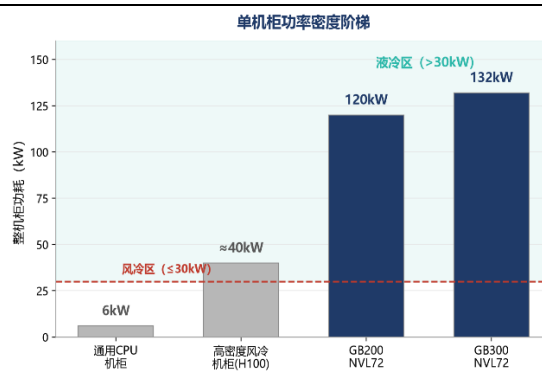
算力密度提升倒逼散热革命，液冷技术正在从“可选项”转变为支撑算力持续发展的“必选项”。AI 大模型的复杂化推动了 GPU、ASIC 等 AI 芯片的算力飞跃，单芯片功耗（TDP）随之持续攀升。英伟达 GPU 功耗从 H100 的 700W，发展到 B200 的近 1000W，再到 GB200 的 1200W，直至 Vera Rubin（VR200）的 2300W，单机柜功率密度也从传统数据中心的 4-6kW，提升至 AI 服务器机柜的 100kW 量级，传统风冷散热极限约在 30kW/柜，已触及物理极限，无法承接如此高的热流密度，叠加全球范围内数据中心能效监管日趋严格，风冷数据中心已无法满足最新 PUE（电能使用效率）要求，液冷成为唯一可行的技术路径。

图1：从英伟达 H100 到 B300，单卡功耗翻倍增长



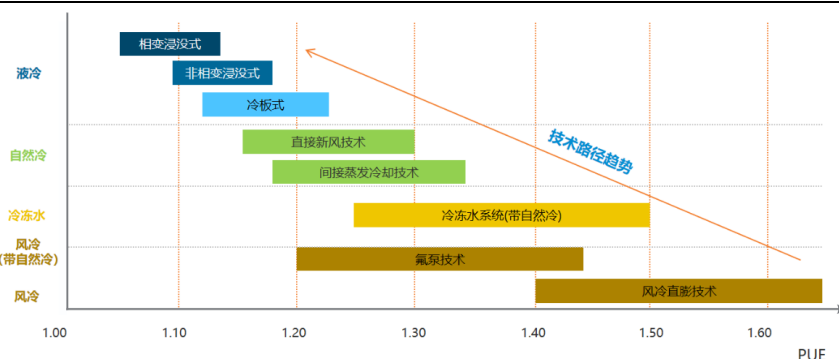
数据来源：Nvidia、Tom's Hardware、开源证券研究所

图2：传统风冷散热极限约在 30kW/柜，已触及物理极限



数据来源：ZTE、Vertiv、Supermicro、HPE、开源证券研究所

图3：风冷数据中心已无法满足最新 PUE 要求，液冷成为唯一可行的技术路径

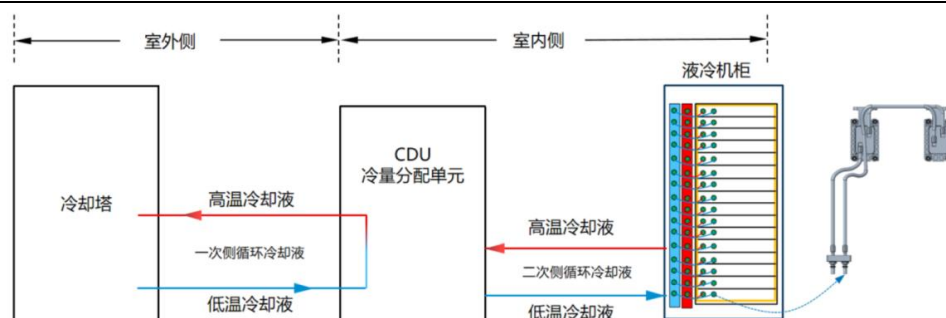


资料来源：《中兴通讯液冷技术白皮书》

冷板式（Direct-to-Chip）凭借成熟供应链与低改造成本快速占领主流市场，浸没式则以极致散热能力和低 PUE 吸引着超高密度场景的目光。按冷却液与发热器件的接触方式，液冷技术可分为直接接触式与间接接触式。直接接触式是指将冷却液体与发热器件直接接触的一种液冷实现方式，涵盖浸没式、喷淋式，散热功率上限更高、潜力更强；间接接触式以冷板式为代表，通过冷板贴合 CPU、GPU 等核心热源传导热量，冷却液不接触电子器件。当前 AI 服务器液冷仍以单相冷板式为主流，其无需改变服务器形态，兼容现有硬件架构与运维体系，供应链成熟度高，可通过 CDU、集分水器、QDC 接头等实现规模化部署，是高功率 AI 机柜的主流散热方案。

冷板式液冷采用微通道强化换热技术具有极高的散热性能，目前行业成熟度较高。冷板式液冷为间接换热机制：冷板与芯片等发热器件紧密贴合，芯片热量通过热传导传递至冷板；在 CDU 循环泵驱动下，低温工质流入冷板流道，经强化对流换热吸收热量后升温；升温后的工质回流至 CDU，通过换热器将热量传递给一次侧冷却液；冷却后的工质再度进入循环，一次侧冷却液最终通过冷却塔将热量排至大气。冷板式液冷系统主要由冷却塔、CDU、一次侧/二次侧液冷管路、冷却介质、液冷机柜组成；其中液冷机柜内包含液冷板、设备内液冷管路、流体连接器、分液器等。冷板式液冷研究积累深厚，技术成熟度较高，是解决大功耗设备部署、提升能效、降低制冷运行费用、降低 TCO 的有效应用方案。

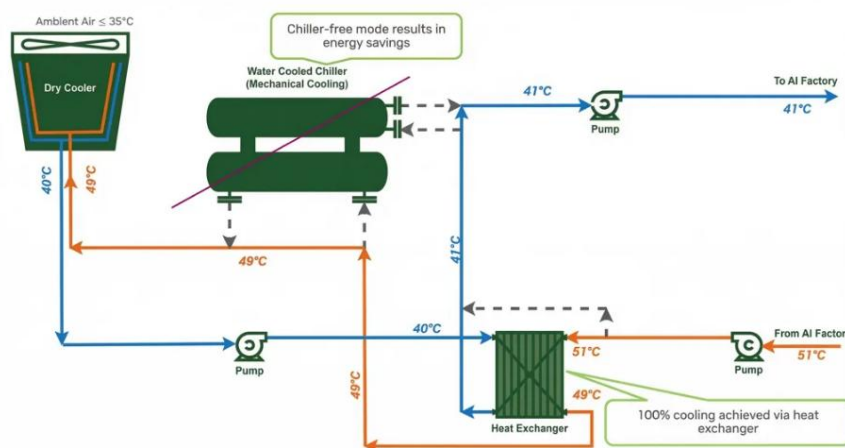
图4：冷板式液冷采用微通道强化换热技术具有极高的散热性能



资料来源：《中兴通讯液冷技术白皮书》

英伟达 Vera Rubin 平台延续 Blackwell 时代冷板式液冷技术路线，在此基础上将全液冷升级为整机标配。官方资料显示，Vera Rubin NVL72 采用单相温水直接液冷方案，最高支持 45°C 供液温度，可依托环境空气自然冷却冷却水，显著降低对冷水机组的依赖与制冷能耗。整机实现 100% 全液冷，第三代 MGX 机架搭载全新内部托盘歧管、机架 UQD08 歧管与支持 5000A 的液冷母线；计算托盘采用无线缆、无水管、无风扇的模块化设计，供电、数据与冷却液均通过盲插接口与背板对接，单托盘组装时间由 GB200 的约两小时大幅缩短至约五分钟。冷却液多采用去离子水或 PG25 丙二醇基流体，闭环系统下使用寿命可达 10 年，日常维护量极低。整体来看，Rubin 未改变冷板式液冷基本技术路径，而是通过温水直冷、全液冷覆盖与模块化盲插设计，实现散热能力与部署效率的双重升级。

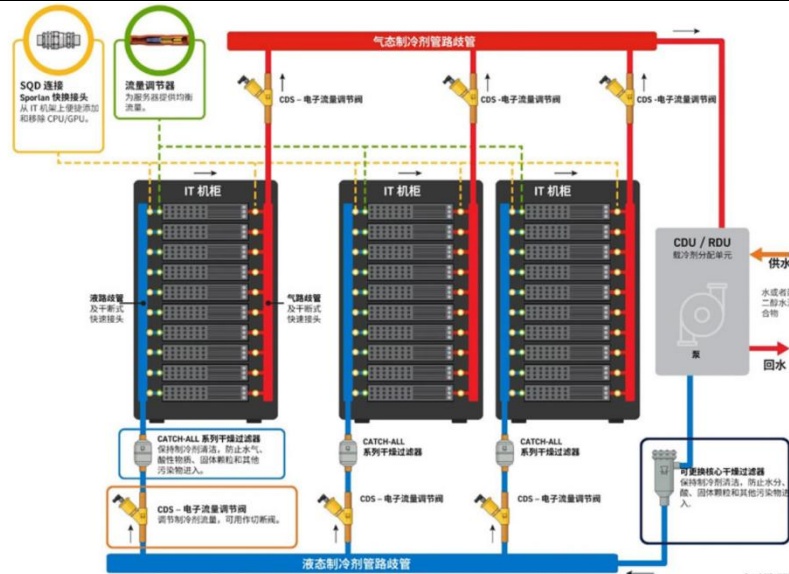
图5：Vera Rubin NVL72 采用 45°C 单相温水直冷，降低冷水机组依赖与制冷能耗



资料来源：Nvidia

面向下一代超高功耗 AI 芯片，两相冷板式液冷有望成为未来重要的发展方向之一。随着芯片功耗与单机机柜功率密度的不断提升，在传统 40℃二次侧供液温度下，单相冷板的散热能力已接近极限。面向未来单芯片超过 3000W 及 MW 级别 AI 机柜的散热需求，若继续采用单相液冷，则需将运行温度降至 30℃以下，这不仅导致自然冷却方式难以适用，还会使单机柜流量与管路尺寸大幅增加，进而影响数据中心液冷系统的节能性、安全性与可部署性。2026 年 3 月英伟达首席热设计专家 Ali Heydari 团队于第 42 届 SEMI-THERM 研讨会发表三篇 IEEE 收录的液冷论文，全部聚焦两相冷板式液冷。该研究为英伟达官方底层散热预研，面向下一代超高功耗 AI 芯片，有望成为下一代高功率机柜储备方案。

图6：两相冷板式冷却效率更好，但其对系统元件介质兼容性、管道承压等要求更高



资料来源：信德迈

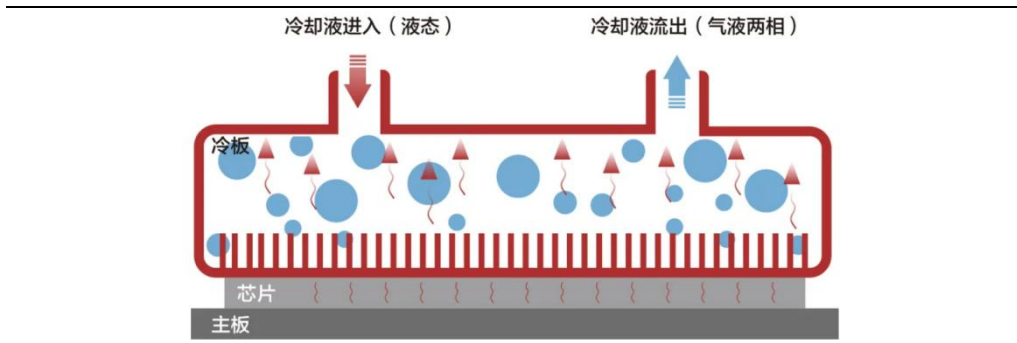
表1：两相冷板式液冷有望成为下一代高功率机柜储备方案

对比维度	单相冷板	两相冷板
传热模式	显热（液体对流）	潜热（沸腾/冷凝）
冷却液状态	始终为液体	液体与气体相互转化
散热效率	较低	高
热流密度上限	100W/cm ²	700-800W/cm ²
应用场景	中等热负荷	高密度、高功率场景

资料来源：新华三、开源证券研究所

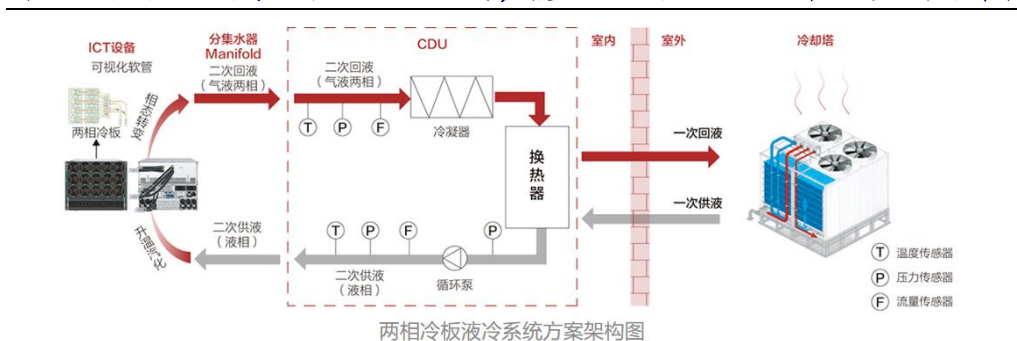
两相冷板技术的本质在于对相变过程的精准调控，其核心原理是借助低沸点冷却工质在密闭腔体内实现“液态-气态-液态”的循环相变，从而达到高效散热的目的。芯片运行时产生的热量经冷板壁面传导进入冷板内部，微通道结构将芯片的高热流密度热量快速传递给工质，使其在冷板内发生相变、沸腾吸热，从而显著提升换热系数并降低器件结温。产生的蒸汽在压差驱动下汇集至回液歧管，随后流向位于冷板远端的 CDU，在板式换热器中与一次侧冷却水换热，释放潜热并凝结为液体；冷凝液温度随之下降。冷凝液在泵驱作用下回流至热源区域，进入下一轮相变循环，由此构成一个封闭的循环散热回路。

图7：液态工质吸热蒸发为气液两相，靠相变潜热带走芯片热量



资料来源：新华三

图8：两相冷板液冷系统采取双回路设计，覆盖 ICT 冷板、CDU 单元与室外冷却塔



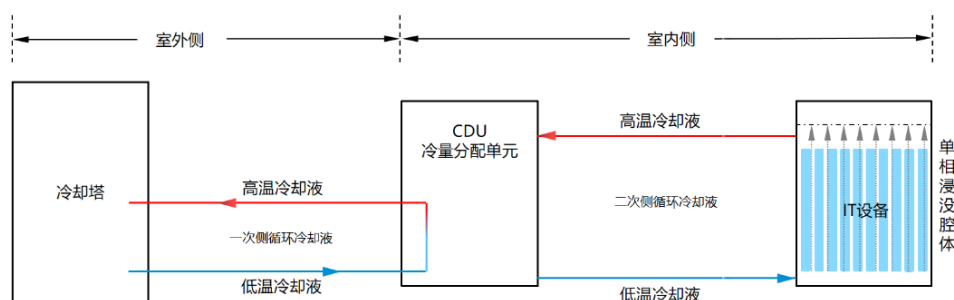
资料来源：新华三

两相冷板液冷技术以相变潜热替代单一显热吸热，兼具高性能与高可用性的双重价值，是当前数据中心散热领域最具潜力的升级路径。相较单相液冷，其散热效率可提升 30%~50%；相较浸没式液冷，则无需整机浸泡，冷却液用量更少、运维更便捷。然而，相变散热对工程可靠性提出更高要求：液体汽化引发的体积膨胀，要求系统具备极高的密封性与压力控制精度，否则易致漏液或气阻，既削弱散热性能又可能损伤电子元件；同时，相变过程动态变化迅疾，温控系统需具备亚毫秒级响应能力，以实时调节冷却状态、抑制芯片温度波动。密封与控制难题，正是当前行业研发突破的核心方向。

浸没式液冷是以冷却液为传热介质，将发热器件完全浸没于冷却液中的制冷形式，发热器件与冷却液直接接触完成热交换。浸没式液冷系统室外侧由冷却塔、一次侧管网与一次侧冷却液构成；室内侧则由 CDU、浸没腔体、IT 设备、二次侧管网与二次侧冷却液构成。使用过程中，IT 设备完全浸没于二次侧冷却液中，故二次侧循环冷却液须采用矿物油、硅油、氟化液等不导电液体。按热交换过程中冷却液是否存在相态变化，可分为单相浸没液冷与两相浸没液冷两类。

在单相浸没式液冷中，作为传热介质的二次侧冷却液在热量传递过程中仅发生温度变化，而不存在相态转变，过程中完全依靠物质的显热变化传递热量。CDU 循环泵驱动二次侧低温冷却液由浸没腔体底部进入，流经竖插在浸没腔体中的 IT 设备时带走发热器件热量；吸收热量升温后的二次侧冷却液由浸没腔体顶部出口流回 CDU；通过 CDU 内部的板式换热器将吸收的热量传递给一次侧冷却液；吸热升温后的一次侧冷却液通过外部冷却装置（如冷却塔）将热量排放到大气环境中，完成整个制冷过程。

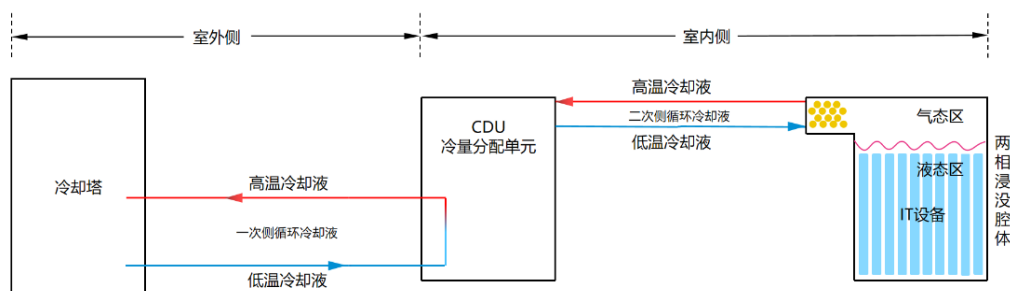
图9：传热介质在热量传递过程中仅发生温度变化，而不存在相态转变



资料来源：《中兴通讯液冷技术白皮书》

在两相浸没液冷中，作为传热介质的二次侧冷却液在热量传递过程中发生相态转变，依靠物质的潜热变化传递热量。其传热路径与单相浸没液冷基本一致。主要差异在于二次侧冷却液仅在浸没腔体内部循环，浸没腔体内顶部为气态区、底部为液态区：IT 设备完全浸没在低沸点的液态冷却液中，液态冷却液吸收设备热量后发生沸腾，汽化产生的高温气态冷却液因密度较小，会逐渐汇聚到浸没腔体顶部，与安装在顶部的冷凝器发生换热后冷凝为低温液态冷却液，随后在重力作用下回流至腔体底部，实现对 IT 设备的散热。

图10：传热介质在热量传递过程中会发生相态转变，依靠物质的潜热变化传递热量



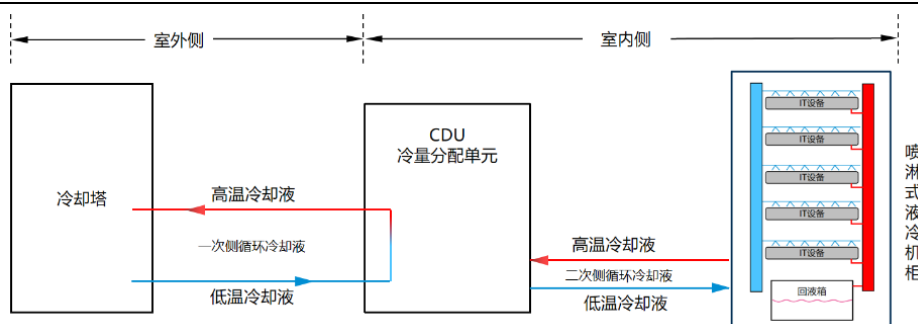
资料来源：《中兴通讯液冷技术白皮书》

浸没式液冷在节能、密度、可靠性与噪声等维度具备显著优势，但其结构颠覆性与工程局限同样突出，是一把典型的“双刃剑”。优势方面：传热路径短、液液换热与汽化潜热换热效率高，无需压缩机制冷，可依托自然冷却与强化通风实现 PUE 低于 1.13；支持单柜散热量达 160kW 的高密部署，机柜无需间隔、机房无需架空地板与冷热通道封闭；设备全浸没于液体，规避温度、振动与灰尘带来的失效风险，配合全液冷免风扇设计实现极致静音。局限方面：浸没腔体采用卧式 Tank，颠覆传统立式机架结构；硬盘、光模块、导热界面材料等器件受冷却液渗入制约需特殊选型；维护需开盖并借助机械吊臂实现设备竖直插拔，复杂度高且伴随冷却液挥发；整柜充液后重量大增，机房地板承重普遍需超过 1500kg/m²。

喷淋式液冷以重力或系统压力驱动冷却液精准喷洒至芯片级发热器件，属于直接接触式液冷，在实现 100%液冷的同时具备优于浸没式液冷的结构兼容性。系统主要由冷却塔、CDU、一次侧与二次侧液冷管路、冷却介质及喷淋式液冷机柜构成，机柜内部集成管路、布液、喷淋模块与回液系统。运行时，冷却液经 CDU 冷却后由泵送至机柜，通过分液器进入与服务器对应的布液装置，或先注入进液箱以获取固定重力势能，再驱动喷淋模块对发热器件及导热元件实施喷淋制冷；吸热后的冷却液经回液箱收集，由泵送回 CDU 完成循环。相较浸没式液冷，其节能效果略逊一筹，

且同样面临冷却液渗入、器件选型受限、维护复杂与机房承重等局限，但结构颠覆性较小、改造成本更低，在高功率密度场景具备一定落地价值。

图11：喷淋式液冷实现了100%液冷，其结构颠覆性优于浸没式液冷



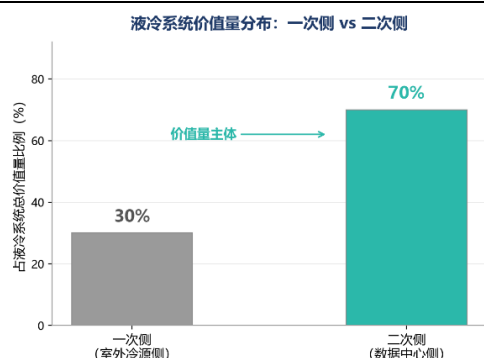
资料来源：《中兴通讯液冷技术白皮书》

从发展趋势看，液冷技术正沿着“冷板式先行、浸没式接力”的主线梯次演进。现阶段，冷板式凭借技术成熟度与对现有服务器架构的良好兼容性占据主流，是存量改造与新建智算中心的现实首选；随着单芯片功耗和单机柜密度的持续攀升，冷板式自身也在向微通道冷板与两相冷板演进，以承接更高热流密度。浸没式凭借散热效率与能效优势，在超高密度场景持续探索，正从技术验证走向规模商用试点；喷淋式则以芯片级精准散热补充特定细分环节。整体来看，各技术路线并非替代关系，而是随功率密度梯度分层适配、长期共存。

1.2、二次侧价值量占比较高，核心部件贡献主要份额

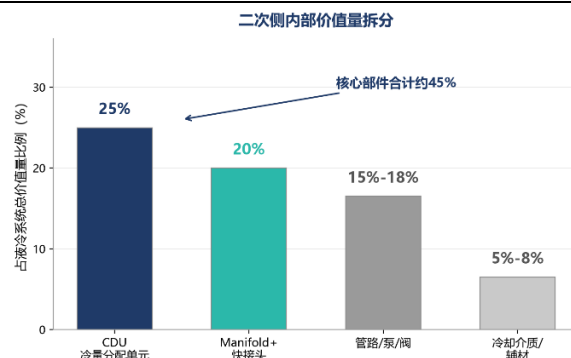
液冷系统由一次侧与二次侧构成，价值量分布主要集中于二次侧。一次侧承担室外排热，价值量占比约30%；二次侧贴近服务器、直接负责芯片热量传导与分配，价值量占比约70%，为液冷价值量最集中的环节。进一步拆解二次侧内部，价值量高度集中于核心部件：CDU（冷量分配单元）集成板式换热器、冗余泵组、阀门与控制系统，价值量占比约25%，是二次侧价值量最高的单一部件；Manifold 分水歧管与快接头合计占比约20%，实现供回液的无工具快速插拔；管路、水泵及阀门占比约15%至18%，冷却介质及辅材占比约5%至8%。整体而言，CDU与Manifold/快接头两大核心部件合计贡献二次侧近半价值量，凸显液冷产业价值向高壁垒核心部件集中的趋势。

图12：液冷系统价值量分布主要集中于二次侧



资料来源：《Global and China Liquid Cooling Technology and Industrial Chain Panorama Industry Research Report》、开源证券研究所

图13：核心部件贡献液冷系统主要价值量份额

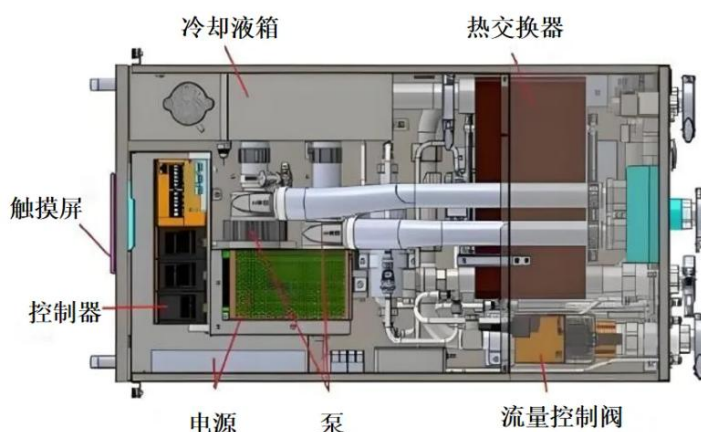


资料来源：《Global and China Liquid Cooling Technology and Industrial Chain Panorama Industry Research Report》、开源证券研究所

（一）CDU：数据中心液冷系统的核心枢纽

CDU（冷量分配单元）是液冷系统中的核心设备，主要为二次侧工质提供冷却和循环动力并对二次侧系统进行集中控制。CDU 是由箱体结构件、热交换器、二次侧水泵、管路组件、配电控制箱及各种传感器组成，其核心作用有四：一是高效搬运热量，驱动冷却液循环，持续将服务器与 GPU 的高密度热量搬出机柜；二是实现物理隔离与系统保护，通过热交换器隔开 IT 侧内循环与冷源侧外循环，规避杂质、压力波动与泄漏风险；三是提供精准动态温控，依托传感器与控制系统按负载实时调节流量与温度；四是支持高密度部署与灵活扩展，推动单柜散热能力跃升至上百千瓦，并可按需扩容，适配不同规模算力建设。

图14：CDU 泵送冷却液冷却服务器，回流至外部冷源散热，形成闭环冷却循环



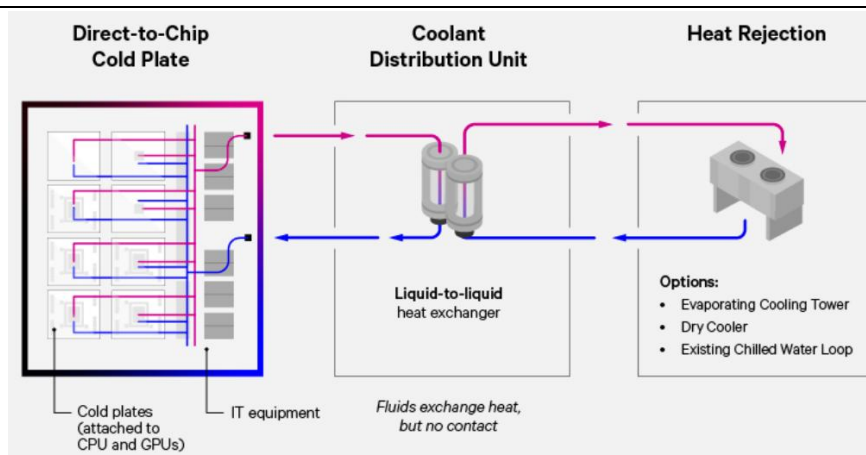
资料来源：艾邦算力资源平台公众号

CDU 作为液冷二次侧的“心脏”与控制枢纽，其竞争壁垒集中于控制算法、可靠性与多架构适配。算法与动态控制上，需在毫秒级响应的流量调节中融入预测与自适应能力，以应对 GPU 集群训练时的短周期热量冲击；可靠性上，面向 7×24 小时不间断运行，须依托工业级元器件、冗余泵阀电源及防漏监测实现无故障切换；架构适配层面，则要求以模块化平台兼容冷板、单相浸没乃至两相浸没等多技术路线。为匹配单柜功率跃升，CDU 正经历全面的能力升级。在换热与泵送环节，换热能力持续扩容，单台换热量不断向更高功率等级突破，以匹配单柜功率的跃升；泵技术由传统机械泵转向无密封电子泵，在降低泄漏风险的同时提升运行可靠性。在可靠性环节，循环配置由 N+1 冗余向 2N 冗余升级，通过关键部件的双重备份，确保单点故障下系统仍可满负荷运行。

（二）冷板：数据中心液冷系统的末端换热核心

冷板是液冷生态系统中真正与芯片物理接触、完成热量转移的关键组件，是液冷系统的末端热交换执行核心。冷板以铜、铝等高导热基材加工成型，内部集成微通道、鳍片及高密度流道结构，是芯片热量接入液冷循环的核心载体。工作过程中，CPU、GPU 等算力芯片产生的高密度热量经 TIM（导热界面材料）传导至冷板基板，有效填充微观间隙、降低界面接触热阻；冷却液以强制对流方式流经内部精密流道，依托大换热面积与强湍流效应高效吸收壁面热量，吸热升温后沿管路回流至 CDU。冷却液在 CDU 内完成换热降温后重新泵入冷板，形成闭环循环，其换热性能直接决定系统的芯片结温控制能力与算力释放上限。

图15: 冷却液经 CDU 送入芯片冷板吸热，回流至散热端冷却循环



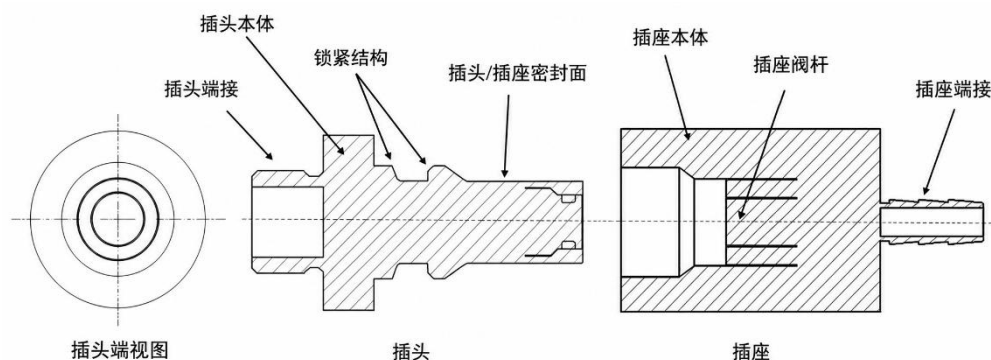
资料来源: Vertiv、开源证券研究所

冷板性能竞争的核心在于热容量、热通量、压力降与热阻四项指标的协同优化。热容量决定冷板的散热总量上限，当前主流产品可覆盖常规芯片散热需求，下一代方案的散热能力持续向上突破，头部先进产品已迈向更高功率层级。热通量衡量单位面积的散热能力，随着 AI 芯片热点集中度持续提升，行业不断向更高流道密度与更优结构方向迭代，散热密度呈持续攀升态势。压力降直接影响泵功耗与系统整体能效，结构更复杂的微通道方案在强化散热能力的同时会抬升流体阻力，行业始终在散热效能与运行功耗之间寻求动态平衡。热阻表征芯片至冷却液的传热效率，热阻越低则芯片结温控制越优、性能释放越充分，行业整体热阻水平随技术迭代持续下探。

(三) QDC: 衔接 CDU 主管路与冷板末端的关键接口部件

QDC (快速接头) 是液冷系统中连接机柜端 Manifold (歧管) 与服务器节点、实现冷却液快速连通与关断的关键部件。QDC 通常由公头与母头配对使用，内置密封件、锁紧机构与导向槽，要求在不使用工具、带液带压的条件下实现不漏滴的快速插拔。其核心作用在于支撑服务器节点与冷却回路的安全热插拔，是液冷系统中最易发生漏液、也最依赖可靠性的连接点。QDC 核心壁垒集中于四点: 密封可靠性，盲插下实现全工况零泄漏，插拔寿命不低于 500 次; 流通性能，有限通径内实现高流量系数、低压力降; 浮动补偿精度，通过多向浮动补偿装配误差; 材料兼容性，长期耐受各类冷却液腐蚀。

图16: QDC 是冷板式液冷和浸没式液冷循环中唯一的动态连接界面



资料来源: Hansen Fluid、开源证券研究所

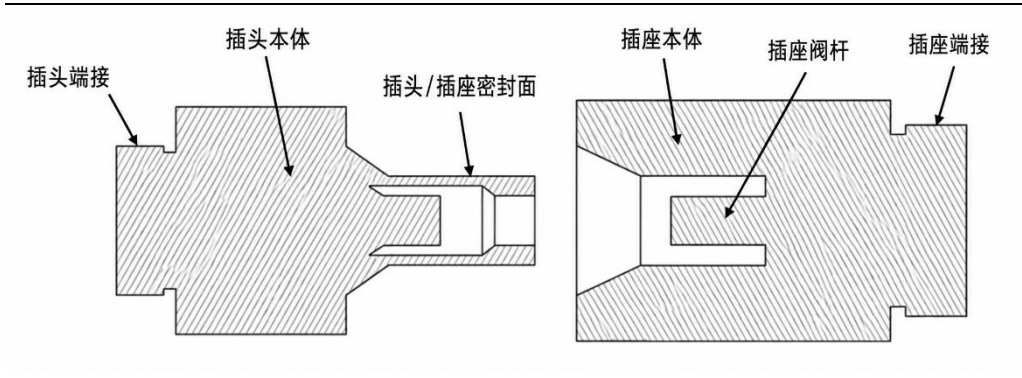
表2：QDC 设计涉及流体力学、材料科学、精密加工及可靠性工程等多个交叉学科

技术指标	行业标准	影响因素	工程意义
通径（DN）	02, 04, 06, 08, 12mm	系统流量需求	决定了单一节点的散热上限
流阻系数（Cv/Kv）	越高越好	内部阀芯结构、流道平整度	降低泵送功耗，提升系统能效
工作压力（PS）	10-16 bar（典型）	材料强度、锁定机构	确保系统在压力波动下不爆管
滴漏量（Spillage）	≤0.2 ml/次	平整端面设计、密封响应速度	保护电子元器件免受液体浸染
容许偏移量	±1.0 mm（径向）	浮动机构设计	降低对机柜装配精度的要求
插拔寿命	500-5000 次	涂层硬度、弹簧疲劳极限	决定了服务器全生命周期的维护频率

数据来源：云帆热管理公众号、开源证券研究所

快速接头的演进正由单一“通断零件”向高集成、智能化的系统组件升级。在插拔方式上，具备高容差补偿能力的盲插连接器（如 UQDB）正逐步取代手动接头，并向液、电、信号一体化盲插演进，实现服务器推入即同步接通；在流道设计上，AI 芯片瞬时功耗攀升推动冷却液流速大幅提升，传统 1/4 英寸接头压降过大，倒逼行业转向大通径设计，并催生电解抛光等流道“微整形”工艺以削减湍流损失；在功能属性上，快插接头通过集成 RFID 或微型压力传感器，实时感知插拔次数与密封老化程度，推动运维由被动维修向主动防御转变；在材料体系上，PFAS 监管收紧促使密封材料向无氟高性能弹性体替代，成为厂商进入欧美市场的准入门槛。整体看，QDC 竞争已从密封与通流等基础指标，延伸至集成度、智能化与材料合规性的多维较量。

图17：UQDC 为盲插型 QDC，具备高容差补偿，支持盲插，适配高密机架运维



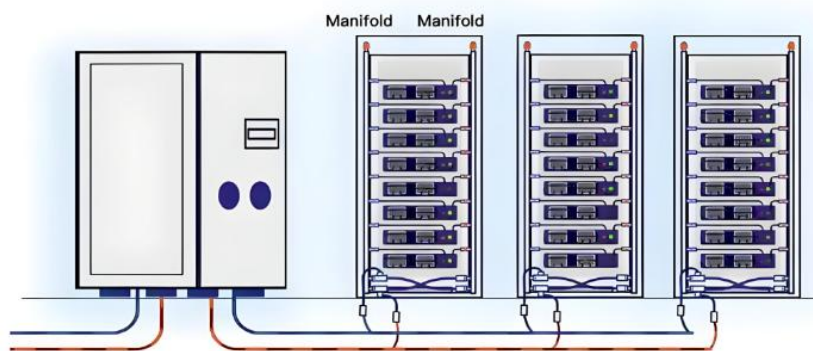
资料来源：Thermax Fluid、开源证券研究所

（四）Manifold：承接 CDU 主管路、均衡支路流量的集配枢纽

Manifold（供回液歧管）是液冷机柜内连接 CDU 与各服务器冷板的核心分配部件，承担着冷却液“分流—汇流”的中枢职能。在结构上，其主流规格为 42U/48U/56U，截面约 40mm×40mm，长度 2.6~3m，材质多为不锈钢 304/316L，具备超强耐腐蚀与高强度特性；在接口形态上分为单排管（适配 10~30kW 机架、接口不超过 12 个）与双排管（适配 50~120kW 机架、接口不低于 16 个）两种，并通过“等长设计”确保各支路流量均匀。冷却液经 CDU 泵送进入供液主管后，通过若干分支节点分流至连

接各服务器冷板的管路。待冷却液在冷板完成吸热后，再经对应节点回流并汇入回液主管，最终返回 CDU，由此形成完整的闭式循环。这一结构如同液冷回路的分流-汇流中心，确保冷却液按需精准分配至各冷板，同时高效汇集吸热后的流体，保障整个回路的有序运行。

图18: Manifold 是连接液冷机柜 CDU 与冷板之间主管路的关键部件



资料来源：云帆热管理公众号

Manifold 作为数据中心液冷系统的关键组件，正沿更高精度、更高效率、更智能化的方向加速演进。在精度与效率层面，行业通过引入先进传感器与控制系统，实现对冷却工质分配的实时监测与智能调控；通过采用新型材料与制造工艺，提升歧管的耐腐蚀性与强度。在智能化与标准化层面，模块化设计与标准化接口的推进，不断增强产品的兼容性与可扩展性，并带动 Manifold、快速接头等关键零部件的解耦交付，使不同厂商产品得以互换兼容。以英伟达 Rubin 架构为例，其配套 Manifold 相较 GB300 已有显著升级：管径进一步加大，互锁球阀由一个增至两个、分别配置于顶部与底部，且不再通过软管连接、直接安装于歧管之上，以适应更高功率密度下的流量与安全冗余需求。

图19: 英伟达 Rubin 高密机架定制升级液冷歧管



资料来源：国际 AI 算力液冷散热技术展

图20: 英伟达 MGX 44RU 机架通用标准液冷歧管



资料来源：国际 AI 算力液冷散热技术展

(五) 冷却液：吸收芯片热量、循环传输冷量的传热介质

冷却液是液冷系统完成热量传递的核心工质，其选型须依据是否发生相变及具体应用场景，并满足多维度性能要求。按是否发生相变，可分为单相与双相两类：单相冷却液在循环中始终保持液态，仅靠温升带走热量，系统结构简单、兼容性高；双相冷却液则借助沸腾与冷凝的相态变化传热，散热效率更高但架构更复杂。按应用场景，冷板式多采用去离子水或乙二醇/丙二醇水溶液，前者导热系数高、成本低，

后者防冻性能优异、适用于室外低温环境；浸没式则涵盖矿物油、硅油、合成油及氟化液等，其中氟化液导热系数与绝缘性俱佳、化学惰性强，但成本较高；高功率双相场景则倾向于低沸点氟化液与碳氢化合物。理想冷却液应同时具备高比热容、高导热系数、防腐防垢等特性，以保障冷却系统的长期稳定运行。

表3：冷却液是液冷系统完成热量传递的核心工质，其选型须依据具体应用场景，并满足多维度性能要求

冷却液类型	技术路径	特点	缺点
去离子水（DI 水）	单相（冷板式）	导热系数高、成本低	需添加缓蚀剂/阻垢剂/杀生剂，并定期更换过滤器
乙二醇/丙二醇水溶液	单相（冷板式）	防冻性能优异，适合低温环境	导热系数与比热容低于去离子水，换热效率相对有限
矿物油（烃类绝缘油）	单相（浸没式）	绝缘性优异、流动性中等、成本低	导热系数低，不适合高功率密度场景
硅油（聚二甲基硅氧烷）	单相（浸没式）	绝缘性优异、耐高温	导热系数低，不适合高功率密度场景
合成油（PAO/GTL）	单相（浸没式）	流动性好、抗氧化性强、导热系数一般	适合中等功率场景
氟化液（全氟聚醚/全氟氨）	单相（浸没式）	导热系数高、绝缘性佳、化学惰性强	成本较高
低沸点氟化液（全氟烯烃/氢氟醚）	双相（冷板/浸没）	沸点 50~90℃、潜热高、绝缘性佳、化学稳定性高，适配高功率场景	成本较高
碳氢化合物（正戊烷/异辛烷）	双相（冷板/浸没）	沸点 30~80℃、潜热较高、成本低	易燃，需腔体密封+安全设计

资料来源：艾邦智造资讯公众号、开源证券研究所

1.3、智算液冷市场持续扩张，国内产业体系加速成型

受益于 AI 服务器功耗持续抬升、单机柜部署密度不断提高，智算液冷市场正进入持续扩张的发展阶段。算力密度的快速攀升倒逼散热体系的价值定位发生重构，过去仅作为服务器配套的辅助散热系统，如今已不再是边缘配套环节，而是直接影响算力能否稳定满额释放、数据中心电费成本高低、以及集群项目能否按期交付的核心基础设施，其在智算算力体系中的战略地位正在被全行业重新定义。赛迪顾问数据显示，2025 年中国液冷数据中心市场规模达 159.8 亿元，同比增长 45.2%，2026 年全球智算液冷市场规模或将突破千亿元。

AI 服务器液冷产业链呈上、中、下游清晰分层，产业竞争正由单一部件角逐向全栈散热能力升级。上游提供基础材料与核心零部件，是技术壁垒最高、利润最丰厚的环节，液冷板领域的奇宏、双鸿、酷冷至尊凭借与英伟达的深度绑定占据主导，CDU 环节由维谛等海外巨头领跑、英维克与申菱环境快速追赶，UQD 长期由史陶比尔等海外企业主导，冷却液在 3M 停产后为巨化股份、新宙邦等国内厂商打开国产替代窗口。中游负责系统集成与服务器制造，浪潮信息、中科曙光、工业富联、超微电脑等服务器厂商主导一体化方案，英维克、高澜股份等专业温控商以全链条能力突围，美的、格力等传统空调厂商亦跨界入局。下游需求主体为谷歌、亚马逊、微软、Meta 等云服务商与字节、阿里、腾讯等互联网巨头，以及移动、电信、联通等运营商。产业链竞争正由单一部件向全栈散热能力升级。

表4：AI 服务器液冷产业链呈上、中、下游清晰分层，产业竞争正由单一部件角逐向全栈散热能力升级

产业链环节	产业定位	具体环节	代表企业
上游	材料与核心零部件	液冷板	奇宏科技、双鸿科技、酷冷至尊等
		CDU	维谛技术、英维克、申菱环境等
		UQD	史陶比尔等

产业链环节	产业定位	具体环节	代表企业
中游	系统集成与服务器制造	冷却液	3M、巨化股份、新宙邦等
		服务器 OEM/ODM	浪潮信息、中科曙光、工业富联、超微电脑等
		专业液冷解决方案商	英维克、高澜股份等
		跨界厂商	美的集团、格力电器等
下游	应用与需求	CSP 厂商	谷歌、亚马逊、微软、Meta 等
		互联网企业	字节跳动、阿里巴巴、腾讯控股等
		电信运营商	中国移动、中国电信、中国联通等

资料来源：AIDC 博寒·液冷展、前瞻网、兴园化工园区研究院、飞鲸投研、制冷快报、LCIE 液博会·2026 第二届国际液冷产业大会暨展览会（北京）、美通社、开源证券研究所

2、计算机产业动态更新

2.1、CSP 厂商与大模型厂商最新动向

（一）OpenAI 发布 GPT-6 Astra，多项基准近满分

9 月 3 日，OpenAI 正式发布 GPT-6 Astra，称其为公司迄今“最智能、最对齐”的模型，目前先向少量机构开放，ChatGPT Plus、Pro、Business、Enterprise 用户及 API、AWS 客户将在随后几天陆续获得访问权限。Astra 最突出特点是能接住复杂目标，将中间几十个步骤一次性做完，可操作浏览器、阅读文件、编写运行代码、使用专业软件并检查结果。在 OSWorld 2.0 测试中 Astra 得 72.6%，高于 GPT5.6 Sol 的 65.7%，完成单项任务平均时间缩短约 47%。

Astra 在数学与抽象推理测试表现突出，FrontierMath Tier 4 取得 97.6%，ARC-AGI-3 达 99.9%，上一代仅 7.8%；在 Terminal-Bench 4.0 取得 57.7%。模型拥有 105 万个 token 上下文窗口，单次最多输出 12.8 万个 token，并加入异步工具调用。定价方面，Astra API 每百万输入 token 收费 10 美元、每百万输出 token 收费 50 美元，均为 GPT5.6 Sol 的 2.5 倍。安全层面，Astra 是 OpenAI 首个达到 Critical 网络安全能力门槛并进入广泛部署阶段的模型，普通版本将拒绝生成高阶漏洞利用代码，所有使用工具的推理过程亦将接受额外监控。

（二）Anthropic 发 Claude 5.1，缓存读取成本降 75%

Anthropic 推出新一代模型 Claude Fable 5.1 与 Claude Mythos 5.1。官方称 Fable 5.1 是目前最强大的公开可用模型之一，专为复杂推理、长周期任务与 Agent 工作流打造，Mythos 5.1 则代表更高能力边界。二者共享相同基础模型能力，但针对不同使用场景采取不同安全措施，其中 Fable 5.1 面向普通用户、开发者与企业开放，Mythos 5.1 面向经审核的高风险领域合作伙伴，保持更严格访问控制。Fable 5.1 多项基准测试表现均优于上一代旗舰 Fable 5。

性能方面，Fable 5.1 在科学研究型 Agent 基准 Terminal-Bench-Science 0.1 中取得 52.6%，较 Fable 5 的 24.7%提升逾一倍；在 Terminal-Bench 4.0 中达 55.8%，Mythos 5.1 进一步达 60.9%。定价方面，输入与输出价格保持不变，仍为每百万 Token 10 美元与 50 美元，缓存读取价格下调 75%，典型任务整体费用较 Fable 5 降低约 25%，高 Agent 化任务成本最高降低约 45%。Mythos 5.1 设计的蛋白质结合剂亲和力达到

Adaptiv Bio 相关竞赛最佳方案的 10 倍，有效结合剂命中率接近 50%。

（三）谷歌发布 Gemini 3.8 Flash，六周内第三款 Flash

谷歌正式推出 Gemini 3.8 Flash，这是其在六周内推出的第三款 Flash 系列模型。该模型在保持与 Gemini 3.7 Flash 相同速度与成本优势的基础上，进一步提升推理与代码能力，重点强化软件工程、Agent 任务及复杂多步骤推理场景。Gemini 3.8 Flash 采用与 3.7 Flash 相同的首发价格，输入每百万 Token 0.75 美元、输出每百万 Token 3.75 美元。同时，谷歌推出面向网络安全领域的专用模型 Gemini 3.8 Flash Cyber，重点提升漏洞发现与自动化补丁修复能力，将通过 Fairwind Program 向经认证的安全防御人员开放。

性能方面，Gemini 3.8 Flash 在 DeepSWE v1.1 长周期软件工程任务中，可自主完成复杂工程问题的端到端解决，性能超过多数参数规模更大的前沿模型；在 HLE-Verified 测试中取得 54.9% 的成绩。其性能提升源于投入更多计算量，模型会执行更多推理步骤并反复调用工具。Gemini 3.8 Flash Cyber 在 CyberGym 基准中展现前沿水平的自主漏洞发现能力，在覆盖 20 种编程语言的内部基准测试中漏洞发现成功率超过 70%，在 CWE-Bench 中 Pass@1 成功率达 47.2%。谷歌已将其用于保护内部代码安全，Chrome 安全团队发现其生成的正确修复补丁数量提升 2.6 倍。

（四）Meta 旗舰 Muse Spark 1.3 发布，性价比反超 Gemini

9 月 3 日，Meta 发布新一代旗舰模型 Muse Spark 1.3，在 Artificial Analysis 的 AI 分析指数榜上以 62 分位列前茅，仅次于 Claude Fable 5.1 与 Claude Opus 5，并反超谷歌同期发布的 Gemini 3.8 Flash（59 分）。Meta 首席 AI 官汪滔称该模型成本低到几乎可忽略不计，并公开向谷歌发难。过去 5 个月 Meta 接连推出四款 Muse Spark 模型，迭代节奏愈发密集。从基准测试看，Muse Spark 1.3 拿下 5 项第一，在长上下文、编程的多数测试优于 GPT-5.6 Sol 与 Claude Opus 5，仅 Terminal-Bench 与 GPT-5.6 Sol 打平。

价格方面，Muse Spark 1.3 的 API 定价与前代 1.2 一致，每百万 token 输入（缓存未命中）1.25 美元、输出 4.25 美元，略高于 Gemini 3.8 Flash，但比 DeepSeek-V4-Pro 高峰期价格更低。开发者实测反馈速度快、成本低，但生成质量一般，在 MineBench 中首测 Elo 评分为 1787 分，较 Gemini 3.8 Flash 的 1888 分仍有差距。Meta 称其优势在于支撑长周期任务与编程，工具调用次数减少 20%、使用 token 数量减少 25%，并增强了对抗鲁棒性与对提示注入攻击的抵御能力。Meta 已规划更大参数规模模型及 Muse Spark 开源权重版本等路线图。

（五）WorkBuddy 开放平台上线，全面开放 Agent 基座

9 月 2 日，WorkBuddy 开放平台正式发布上线，首批引入超百家生态伙伴，面向智能硬件、行业应用与开发者全面开放底层 Agent 能力，系国内 AI Agent 赛道首个同时打通硬件、应用、开发者三层生态的开放平台。腾讯云副总裁、腾讯 CodeBuddy 与 WorkBuddy 负责人刘毅表示，希望把 WorkBuddy 打造成面向 Agent 时代的操作系统，做一个能承载所有工具的平台。平台已接入智能眼镜、录音卡片、耳机等十余个品类、超 30 个品牌，首发 9 款联名硬件。

Buddy 应用同步上线，伙伴可基于 WorkBuddy Agent 底座能力构建品牌化 AI 工作台，无需自研 AI，目前已覆盖金融、SaaS、法律、医疗、教育等 20 多个领域，由通达信、广发证券、腾讯 SSV、腾讯健康、北大法宝等 30 多个伙伴共建。平台同步

开放 Skill、Expert、Connector 三大能力，调用标准 API 即可获得完整 Agent 能力，Expert 市场涵盖 Agent 型与 Team 型两种形态，Connector 支持 MCP 与 CLI 双方案。自今年 3 月初上线至今，WorkBuddy 已完成 50 多个版本迭代，在政务、教育、零售、金融等 50 多个行业加速落地。

（六）DeepSeek-V4 多模态模型开源，登顶多项榜单

8 月 31 日，DeepSeek-V4-Flash-Vision-Exp 正式开源，这是 DeepSeek-V4 系列中的首个实验性多模态模型，此前曾于 8 月 21 日上线 DeepSeek API 平台。该模型基于 DeepSeek-V4-Flash 架构，通过集成视觉模块并经过持续训练解锁视觉理解能力。基准测试显示，其多模态智能体能力进一步提升，在 7 项纯文本智能体任务中取得 5 胜 1 平 1 负；在真实世界软件工程测试 DeepSWE 中以 59.3% 的成绩反超 Opus-4.8，成为第一名；在零样本视觉推理测试 ZeroBench 及 Agents' Last Exam 综合智能体测试中击败 Opus-4.8，登顶榜首。

在工具调用能力上，该模型在 Toolathlon-Verified 中得分 75.9%，仅落后 Opus-4.8 约 0.3 个百分点；但在自然语言转代码仓库测试 NL2Repo 和数据科学测试 DSbench-Hard 上，与 Opus-4.8 仍有约 10% 差距，显示其在高度复杂、结构化的数据科学任务上仍有优化空间。过去一个月内，DeepSeek-V4 系列已完成多轮更新，其中 7 月 31 日 DeepSeek-V4-Flash 正式版 API 上线公测，8 月 13 日 DeepSeek-V4-Pro 正式版 API 发布，8 月 19 日 DeepSeek Harness 迭代至 v0.1.0-rc.8 版本。

（七）腾讯混元发布 Hy4 preview，开源并推出轻量版

8 月 28 日，腾讯混元发布 Hy4 preview，总参数 770B、激活参数 49B、上下文长度 1M，在代码、办公、科学等真实生产力任务上展现出出色能力。该模型通过与腾讯内部软工、游戏、金融、安全等领域专家高质量数据共建，在软件工程、办公分析、游戏开发与科学研究四类场景取得显著进步。盲测显示，Hy4 preview 均分 2.99/4.00，略优于 GLM-5.3（2.92/4.00）与 Kimi K3（2.94/4.00）。

在科学研究侧，Hy4 preview 自主分析推理系统瓶颈，端到端吞吐较基线提升 31.8%；配合 Hyra 在机器学习力场分子动力学模拟上进一步提速 2.0 倍，单张高端 GPU 可容纳 30 万原子；在三维 Blaschke-Lebesgue 几何难题上将体积下界从 0.380799 推进至 0.41104，距最终证明仅剩约 2% 差距。模型已开源，在腾讯云 TokenHub 与 OpenRouter 上线，并可在 WorkBuddy、CodeBuddy、元宝、ima 等产品体验。

由于模型权重接近 1.5TB，为满足配置有限机器的运行需求，9 月 1 日腾讯混元团队推出轻量版，通过 Sherry 稀疏三值量化算法与 MIX-STQ1_0 混合精度策略，将权重压缩至约 214GB。Sherry 以四个权重为一组量化成三值并强制每组留一个零，平均 1.25 比特；MIX-STQ1_0 按敏感度逐层分档，整体量化误差更低。压缩后模型在主流评测集上与原版差距多在一两分以内，并领先 UD-IQ1_M 量化版本。腾讯混元还与 prima.cpp 验证了异构设备联合推理，将 214GB 模型协同跑至 1.02 token/s，较笔记本单机 offload 快约 6 倍。

2.2、国产算力产业最新进展

（一）长鑫科技 LPDDR6 全球首发量产，首搭小米 18 Fold

8 月 29 日，长鑫科技宣布自研低功耗内存 LPDDR6 实现量产，首批搭载于小米 18 Fold 折叠旗舰手机。雷军微博公开祝贺长鑫实现 LPDDR6 量产，称未来将有更多

中国科技企业强强联合、共同攀登技术高峰。此次官宣距离其研发验证进入尾声不足一个月。

长鑫率先量产落地 LPDDR6，实现全球首发量产，中国存储企业首次在高端内存标准上打破海外厂商长期垄断先发的局面。据半年报及供应链消息，长鑫 LPDDR6 峰值速率 12800Mbps，与 SoC 协同基础速率 10667Mbps，最高容量 16GB，与三星、SK 海力士路线图初代产品规格同档位。小米产业主体已完成投资，实现供应链与股权双重绑定；苹果亦有望与长鑫在存储供应上合作，长鑫内存计划仅供国行入门版 iPhone、iPad，受管制规则限制苹果只能采购标准现货。

（二）华虹宏力获大基金三期增资，斥 280 亿扩 12 英寸产线

9 月 1 日，华虹宏力发布对外投资公告，公司及全资子公司联合三家机构对无锡华虹宏力共同增资 41.70 亿美元，约合人民币 280 亿元，用于建设产能约 5.5 万片每月的 12 英寸特色工艺晶圆代工产线。华虹宏力称，此举是为把握半导体行业发展机遇，优化产业布局、适度扩大产能规模以增强竞争优势。其中，华虹宏力及全资子公司合计出资约 143 亿元，另三方机构合计出资约 137 亿元。

本次增资引入无锡国资委、大基金三期、国开行三家投资方。其中无锡锡虹国芯二期实控人为无锡市国资委，华芯鼎新的大基金三期出资比例达 99.9001%，国开公司实控人为国家开发银行。增资完成后，华虹宏力及全资子公司合计持股 51%并处于控股地位，无锡锡虹国芯二期、华芯鼎新、国开公司分别持股 20%、15%、14%，国开公司另享有回购权及定期收益权两项安排。

（三）液冷赛道订单排至年底，智算液冷市场将破千亿

8 月 30 日，据央视财经，随着万卡智算中心、超节点等高功率算力集群加速落地，芯片功耗不断攀升，传统风冷难以满足散热需求，液冷迅速被推到台前。业内人士表示，单机功率超过 100 千瓦后，冷板式液冷压力会明显增加，当 AI 服务器机柜功率达到 200 千瓦以上时，浸没式方案更有优势。在需求扩张下，竞争传导至生产端，山东济南高端 AI 服务器产线上，一批液冷服务器正加紧装配测试，山东青岛莱西一处液冷核心设备部件 CDU 组装产线订单已排到 12 月底。

广东佛山一家液冷部件生产企业今年的订单、产值与交付量同比均接近 100% 增长。赛迪顾问数据显示，2025 年中国液冷数据中心市场规模达 159.8 亿元，同比增长 45.2%。行业机构预测，2026 年全球智算液冷市场规模或将突破千亿元。只要 AI 服务器功率持续上升、机柜密度持续提高、智算中心持续建设，智算液冷市场将继续扩张。

（四）范式智能超 10 亿元采购华为昇腾 950 芯片

9 月 2 日，据新浪科技独家获悉，港股上市公司范式智能（原第四范式）近期与华为达成大规模算力订单合作，计划出资超 10 亿元采购华为昇腾 950 系列芯片，用于支撑 AI 大模型在金融、制造等核心领域落地及企业 AI 业务扩张。这也使范式智能成为首批将国产最高端算力底座纳入核心生产系统的 AI 企业之一。

围绕此次合作，范式智能创始人兼 CEO 戴文渊专程赴深圳，与华为常务董事杨超斌举行高层战略座谈，双方多位核心管理层共同出席，就算力合作的具体落地进行磋商。新增算力资源将主要用于范式智能在手订单的项目交付及 AI 产品研发。

2.3、海外算力产业最新进展

（一）OpenAI 采购数万台 Mac 做强化学习，Mac 业务增 29%

据 Information 消息，OpenAI 已购买数以万计的 Mac mini 和 Mac Studio，专用于强化学习，Anthropic 亦通过亚马逊云服务租用 Mac mini 承担类似任务。这些 Mac 被用于训练“计算机使用智能体”，即能自主操作电脑、完成编辑测试代码、整理邮箱、总结文档等多步骤任务的 AI 系统。这股热潮直接反映于苹果财报，最近一个季度 Mac 销售额同比增长近 29%，达 103 亿美元，成为苹果增速最快的业务。

Mac 能在强化学习场景被大规模采购，核心在于统一内存，苹果 M 系列芯片采用单一共享内存池，相比英伟达 GPU 显存与系统内存分离的结构具备性能优势；同时 Mac mini 和 Mac Studio 配备专门散热系统，长时间运行 AI 任务不降频。英伟达已将苹果视为本地 AI 领域最大竞争对手，并推出 DGX Spark 直面竞争。苹果方面，高配版 Mac mini 和 Mac Studio 已断货数月，苹果前 AI 产品企业营销经理 Todd Dailey 透露部分企业已转向有现货的 DGX Spark 等替代方案。

（二）英伟达 868 亿元收购 Hugging Face，从算力层迈向生态层

9 月 3 日，英伟达 CEO 黄仁勋宣布，公司已达成协议，以 129.303 亿美元（折合人民币 868 亿元）收购 AI 开源社区平台 Hugging Face。交易完成后，Hugging Face 将加入英伟达体系，双方计划扩大平台规模、加强基础设施建设，并推动更多开发者和机构使用 AI。据英伟达披露，Hugging Face 平台目前拥有超 1800 万名开发者、研究人员与创作者，汇聚超 300 万个模型、50 万个数据集及 100 万个应用，全球已有超 20 万家公司通过该平台将 AI 技术应用于实际业务。

英伟达强调，Hugging Face 仍将保持开放平台定位，开发者依然可自由选择模型、开发框架、云服务商与计算平台，使用 Hugging Face 并不意味着必须采用英伟达 GPU 或基础设施，平台亦将继续支持来自不同公司的开源模型与开放权重模型。黄仁勋称，过去十年 Hugging Face 团队打造了全球最重要的开放模型社区之一。英伟达目前已成为该平台最大的开放模型与数据贡献者之一，累计发布超 500 个模型及超 250 个开放数据集。

此番收购意味着英伟达正从单纯 AI 算力供应商，向覆盖模型、工具、开发者与应用生态的综合 AI 平台进一步扩张。过去几年 GPU 已成为 AI 训练与推理的核心硬件，随着大模型竞争进入应用与生态阶段，模型平台、开发者社区及工具链的重要性不断提升。英伟达表示，将利用自身基础设施、工程能力与全球资源，提升 Hugging Face 在可靠性、安全性、模型评估、推理与部署方面的能力。开源社区的价值建立在开放与多元选择之上，如何平衡商业利益与开放精神，将成为此次交易最大的考验。

（三）三星锁定七成产能至 2031 年，HBM 现货价飙至合同四五倍

9 月 1 日，据韩媒报道，三星存储业务部门已将 2031 年前约 70% 的产能分配给长约订单，主要签约客户包括英伟达、微软及谷歌等大型科技公司。然而即便是这些大型客户也无法取得合同约定的全部供货量，HBM 抢货激烈，进一步推升现货价格。据内存市场研究公司 MegaGrid Supply 数据，36GB 的 HBM3E 产品现货售价为 2100 美元，相当于长期合同价格的四到五倍；16 层 HBM4 产品若经现货市场购买，交易价格高达 3500 美元。

供应压力正从 HBM 蔓延至一般 DRAM，HBM 生产吸收大量 DRAM 产能，且仍处量产初期的 HBM4 良率低于 HBM3E、耗用更多 DRAM，使整体短缺进一步恶化。三星电子与 SK 海力士均认为短缺短期无法解决，正寻求产能扩张。受此影响，价格上涨预计将为两家带来创纪录盈利，市场普遍预期三星第三季度营收 206.64 万亿韩元、营业利润 116.38 万亿韩元，SK 海力士第三季度营收 101.76 万亿韩元、环比增长 28.2%。韩国半导体产业协会执行董事安基铉表示，即便新厂投产也只能缓解而无法彻底消除短缺。

（四）英伟达借租约介入 Anthropic 350 亿美元算力大单

8 月 31 日，据《华尔街日报》消息，Anthropic 已与英伟达支持的云服务商 Lambda 签署 350 亿美元云计算协议，这是其 2026 年第二笔百亿美元级算力大单，两笔合计达 800 亿美元。交易中，英伟达通过持有数据中心租约介入，为 Anthropic 提供信用背书，Lambda 则部署芯片并供应云算力，英伟达角色由此从芯片供应商延伸至基础设施撮合方。

Anthropic 采购紧迫源于产品高速放量，其年化收入从 2025 年底约 90 亿美元增至今年 3 月底突破 300 亿美元。此前其已与云服务商 Nscale 签署 450 亿美元协议，半年多集齐亚马逊、谷歌、英伟达、微软四大巨头算力，承诺超 11 吉瓦。因 Anthropic 无信用评级，英伟达所持租约在交易中充当信用中介，较此前因反垄断担忧暂停的收入分成计划，本次方案既提供信用背书，又规避了客户控制权争议。

（五）三星电机获 1.07 万亿韩元 MLCC 大单，行业迈入涨价循环

9 月 2 日，三星电机宣布与一家全球知名客户签署价值 1.07 万亿韩元（约合人民币 52.54 亿元）的 MLCC 供应合同，用于其 AI 服务器，规模占公司 MLCC 事业部 2025 年营收约 5.2 万亿韩元的 21%。三星电机在全球 MLCC 市场份额超 40%，已与十余家全球企业签署长期供货合同，自 2025 年 5 月以来在 AI 服务器硅电容与 MLCC 领域累计斩获总价值 2.25 万亿韩元合同，正加速调整业务组合，以高附加值产品为核心。

随着 AI 服务器中 GPU、HBM 大量应用，高可靠 MLCC 需求走高，三星电机已率先针对 OEM、ODM 客户调涨 2026 年第四季度报价，其中消费级 X5R 上调 25% 至 30%，AI 服务器高端 X6S 平均涨幅 10% 至 20%。国巨、村田、太阳诱电等亦传出涨价信号。国海证券认为被动元件行业景气度高，明年高容 MLCC 供需缺口或扩大，紧张格局有望延续至 2028 年；TrendForce 预计第四季供需缺口扩大、价格走升，高盛预测全球 AI 服务器 MLCC 市场将以年均约 34% 增长，从 2025 年 14 亿美元增至 2030 年 58 亿美元。

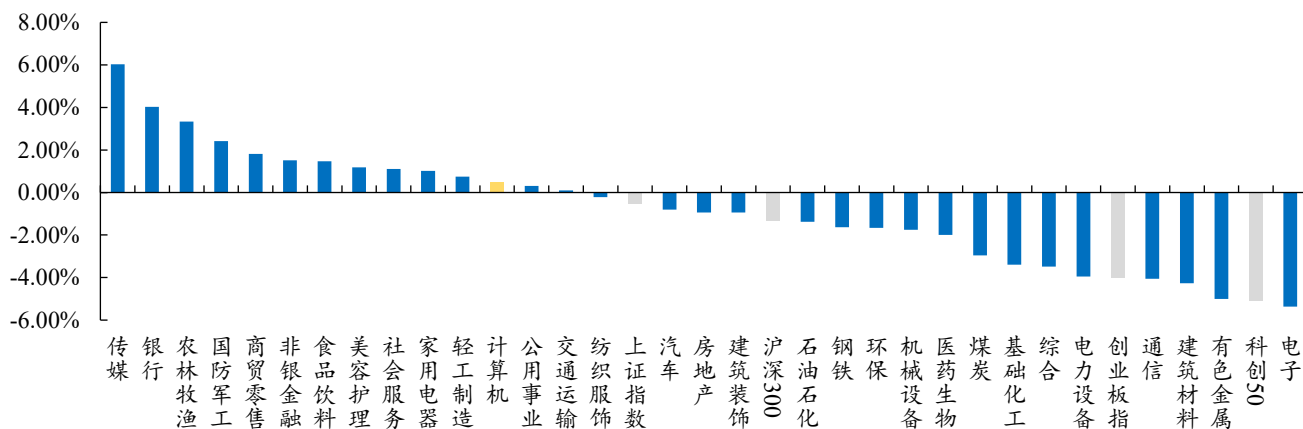
3、上周市场表现回顾（2026/8/31-2026/9/4）

计算机指数上涨 0.49%，相对主要市场指数表现较强。同期沪深 300 指数下跌 1.33%，上证指数下跌 0.56%，创业板指下跌 4.03%，科创 50 指数下跌 5.10%，计算机指数分别跑赢 1.82、1.05、4.52 和 5.59 个百分点。

横向比较一级行业，本周传媒、银行、农林牧渔表现居前，分别上涨 6.03%、4.03% 和 3.34%；计算机板块上涨 0.49%，在 31 个一级行业中排名第 12 位。与此同时，电子、有色金属、建筑材料、通信和电力设备分别下跌 5.36%、5.01%、4.26%、4.06%

和 3.95%。整体来看，在本周成长板块普遍承压的背景下，计算机板块体现出一定的相对韧性。

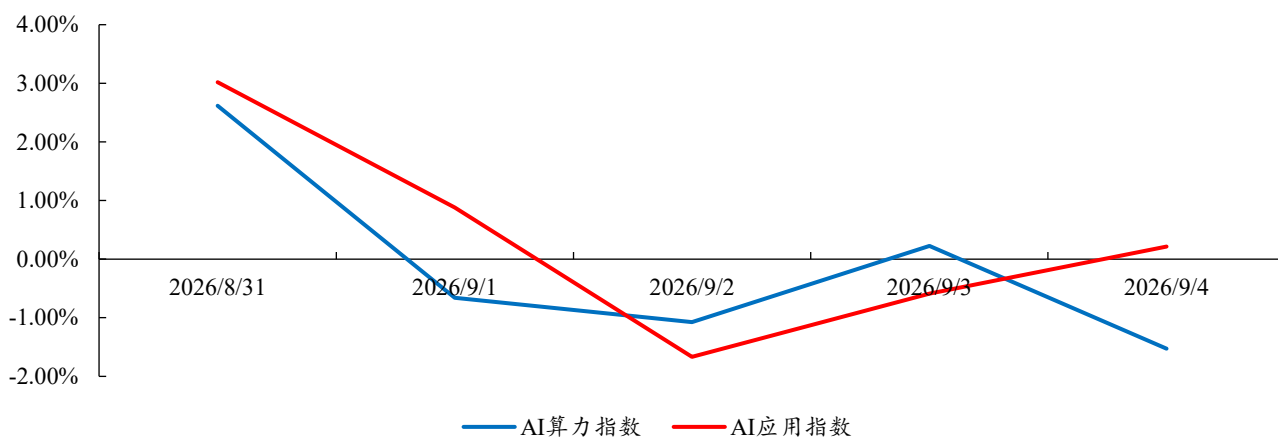
图21：计算机行业板块相对表现



数据来源：wind、开源证券研究所

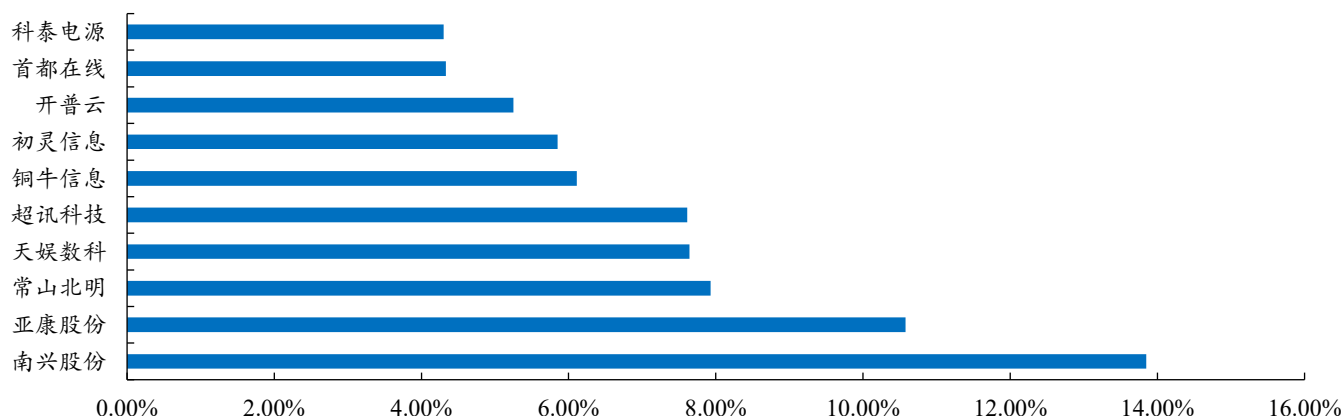
本周 AI 细分方向呈现一定分化，AI 应用指数整体表现明显优于 AI 算力指数。按照日度涨跌幅复合计算，8 月 31 日至 9 月 4 日 AI 应用指数累计上涨约 1.81%，AI 算力指数累计下跌约 0.47%，AI 应用相对 AI 算力取得约 2.28 个百分点的超额收益。

图22：AI 细分指数表现情况



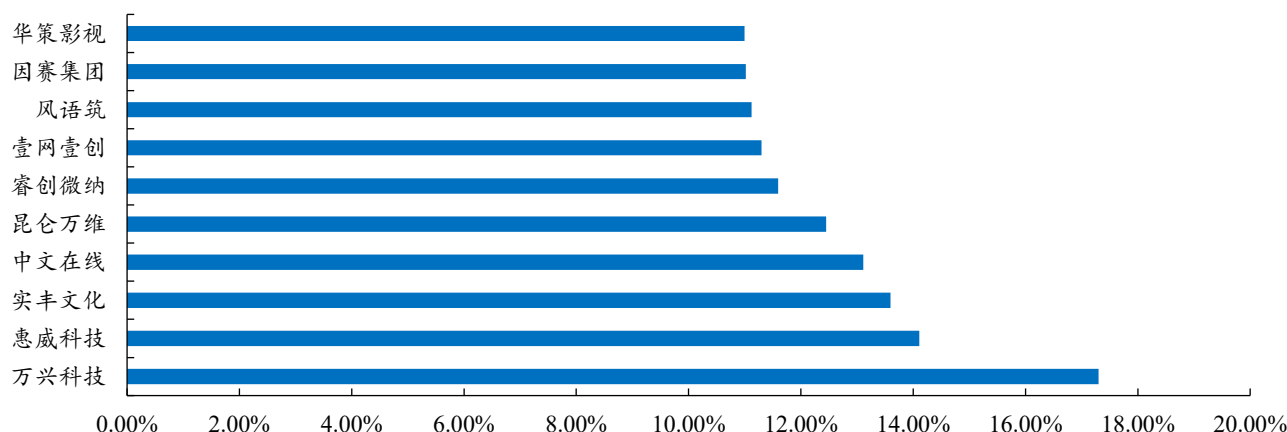
数据来源：wind、开源证券研究所

AI 算力指数整体小幅下跌，但板块内部仍存在较为明显的结构性行情。南兴股份本周上涨 13.85%，亚康股份上涨 10.58%，均录得两位数涨幅。

图23：AI 算力板块个股本周涨幅前十


数据来源：wind、开源证券研究所

AI 应用方向个股表现更为强势，板块赚钱效应更加突出。涨幅前十个股周涨幅全部超过 11%，万兴科技本周上涨 17.30%，位居 AI 应用板块涨幅首位。

图24：AI 应用板块个股本周涨幅前十


数据来源：wind、开源证券研究所

4、风险提示

大模型技术迭代不及预期风险。大模型技术迭代是推动产业能力边界拓展与下游场景深化的核心驱动力，当前行业正处于多模态感知、复杂推理、智能体执行等前沿技术的攻坚突破期，技术突破节奏、能力落地效果均存在一定不确定性。若后续核心技术突破进度、模型能力提升幅度低于市场预期，无法有效支撑下游高价值场景的深度适配与功能升级，可能导致产业落地节奏整体延后，行业成长进程存在阶段性放缓的可能性。

AI 技术应用不及预期风险。下游场景规模化渗透是大模型产业价值兑现的核心支撑，当前大模型正从技术概念验证期向垂直行业规模化应用期过渡，落地进度受行业数字化基础、客户付费意愿、投入产出比验证等多重因素共同制约。若后续垂直行业适配进度、企业付费转化节奏不及市场预期，产业商业价值的兑现周期可能有所拉长，行业长期成长动能的释放节奏存在一定波动可能。

相关政策不及预期风险。大模型产业发展高度依赖政策环境的引导规范与配套支持，当前人工智能领域处于快速发展阶段，数据安全合规、技术监管标准、产业扶持政策、市场准入规则等相关制度仍处于持续完善过程中。若后续行业监管要求趋严、扶持政策落地进度慢于预期，或合规要求抬升导致企业研发与运营成本增加，可能对大模型技术迭代、商业化落地及行业竞争格局形成阶段性扰动，行业发展节奏存在相应调整的可能性。

AI Infra 建设不及预期风险。AI 基础设施是大模型技术迭代与产业规模化落地的底层算力与工程支撑，算力集群建设、高速互联部署、存储体系配套等基建环节具备投入规模大、建设周期长、产业链协同要求高的特征，落地进度受硬件供给、资金投入、工程能力等多重因素约束。若后续高端算力供给受限、基建项目建设进度慢于市场预期，可能形成算力供给瓶颈，制约大模型训练迭代与应用规模化落地，行业整体发展节奏存在阶段性承压的可能。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已正式实施。根据上述规定，开源证券评定此研报的风险等级为R3（中风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C3、C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C3、C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中的具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（Outperform）	预计相对强于市场表现 5%~20%；
	中性（Neutral）	预计相对市场表现在-5%~+5%之间波动；
	减持（Underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（Overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（Underperform）	预计行业弱于整体市场表现。

备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数（北交所基准指数为北证 50 指数）、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海

地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

北京

地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

深圳

地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

西安

地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn